Capstone-1
Supplier Pricing Prediction and Segmentation

Domain Area: Supply Chain | Data Science

Prepared by:
Prashant Sanghal

Motivation behind this Project?

- Caterpillar (manufacturer of construction equipment) procured 8,855 unique tube assemblies of varying specifications from 57 different suppliers.
- Every time, company needed a tube assembly, they had to issue a request for quote to identify best cost qualified supplier.
- This resulted in 30,213 quote requests over 35 years and buying same assembly at multiple price points from various suppliers.

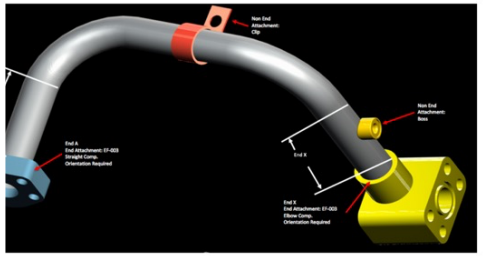
Business Question:

- Can we predict supplier pricing from previously supplied quotes and help business save time and cost?

- Can we identify a group of best managed assemblies and suppliers who fully meet expectations against various business needs?

Equipment

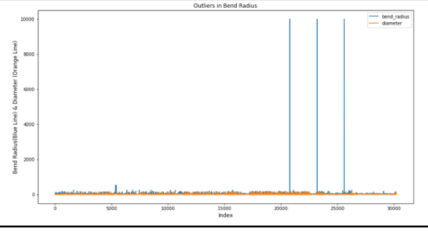
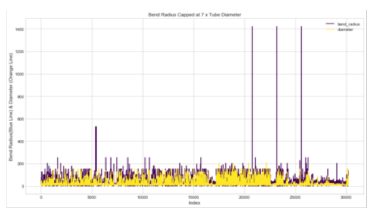
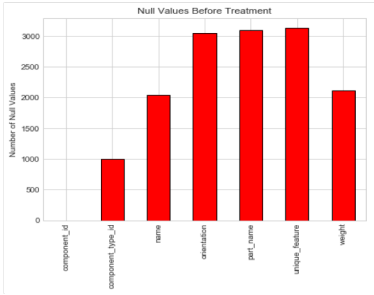
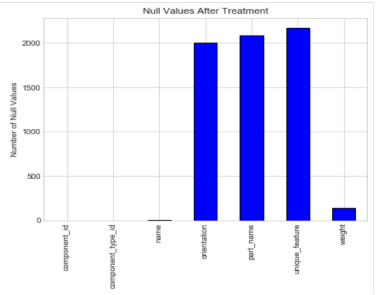
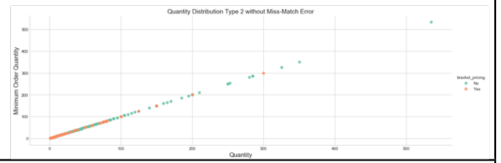
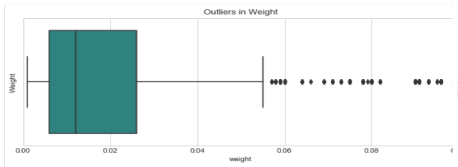
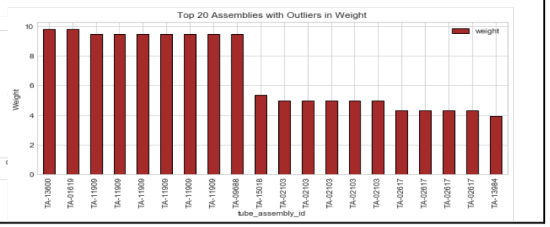
Tube Assembly



Data Collection and Feature Enhancement

- Kaggle Dataset
- 21 tables: tube specifications, order history
- 77% Missing Data
- 1000+ Categorical features
- Contained outliers, cost/unit error and infinity number

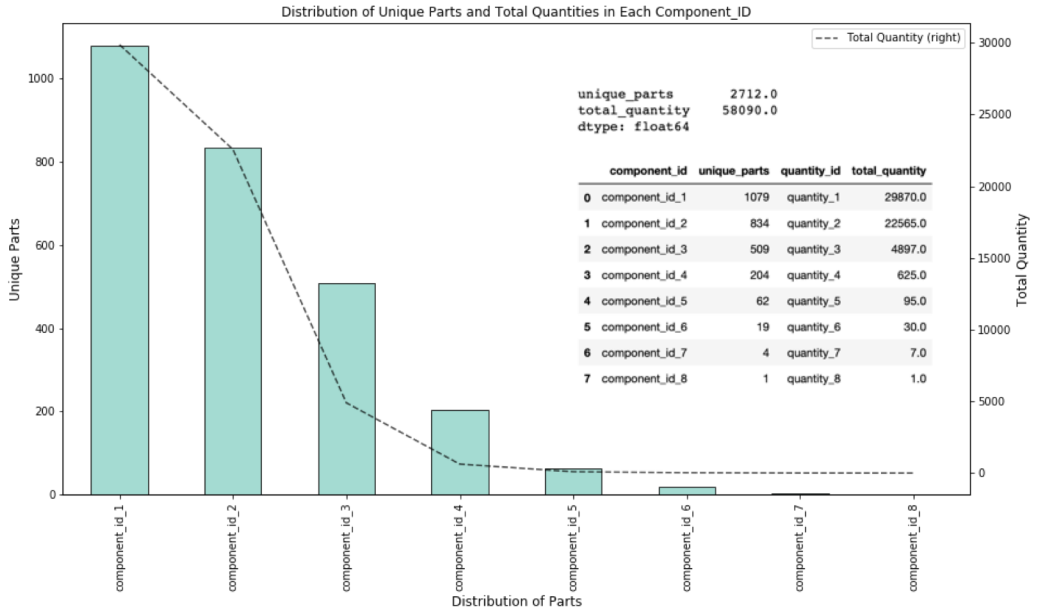
- Feature Engineering:
  - Predictive: Derived from existing features
  - Clustering: Developed from Business needs

Before Data Treatment	After Data Treatment
• Outliers in Bend Radius (Before: 9999)	Outliers in Bend Radius (After Modification: Capped at 1400)
	
• Null values (Before: Range 0 to 3000)	Null Values (After Fill Method: Range 0 to 2500)
	
• Cost per quantity miss-match error (Before-Scattered)	Cost per quantity without miss-match error (After-Linear)
	
• Outliers in Weight (Heavy Assemblies without Treatment)	
	

Exploratory Data Analysis for Actionable Insights

What makes up the assembly ?

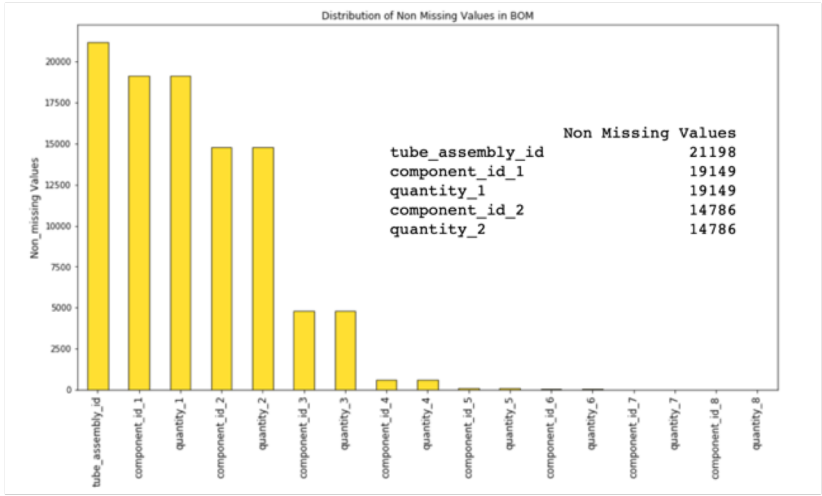
- Breakdown of assembly components




```
unique_parts       2712.0
total_quantity    58090.0
dtype: float64
```

	component_id	unique_parts	quantity_id	total_quantity
0	component_id_1	1079	quantity_1	29870.0
1	component_id_2	834	quantity_2	22565.0
2	component_id_3	509	quantity_3	4897.0
3	component_id_4	204	quantity_4	625.0
4	component_id_5	62	quantity_5	95.0
5	component_id_6	19	quantity_6	30.0
6	component_id_7	4	quantity_7	7.0
7	component_id_8	1	quantity_8	1.0

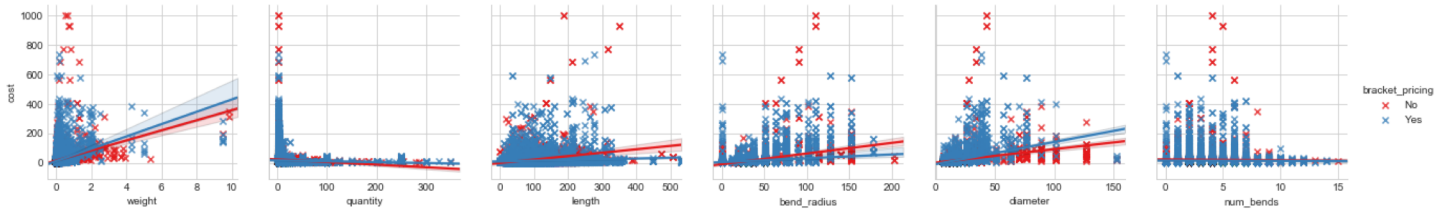
Distribution of non-missing values in BOM




```
                  Non Missing Values
tube_assembly_id                21198
component_id_1                  19149
quantity_1                      19149
component_id_2                  14786
quantity_2                      14786
```

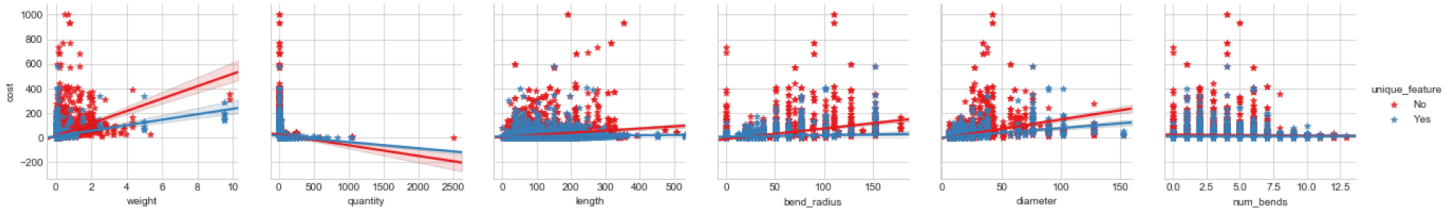
Supplier Cost Dependency:

Cost Vs Tube Specification by Bracket Pricing



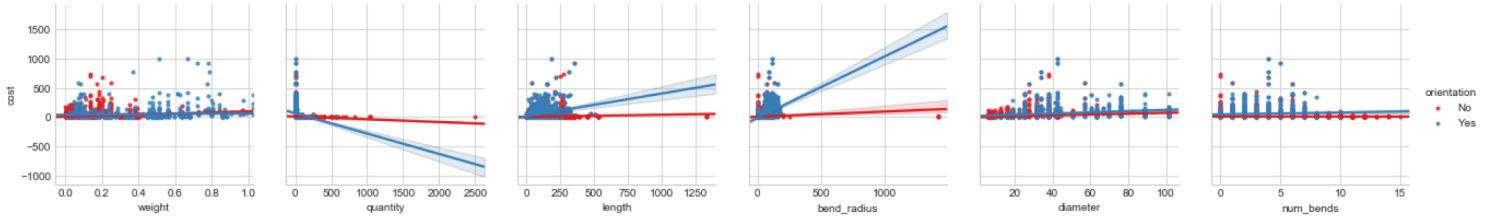
- Cost increases with weight, length & bend radius.
- Cost Decrease with increase in order quantity.

Cost Vs Tube Specifications by Unique Features



- Cost increases by adding new features.
- Assemblies without unique features more costlier

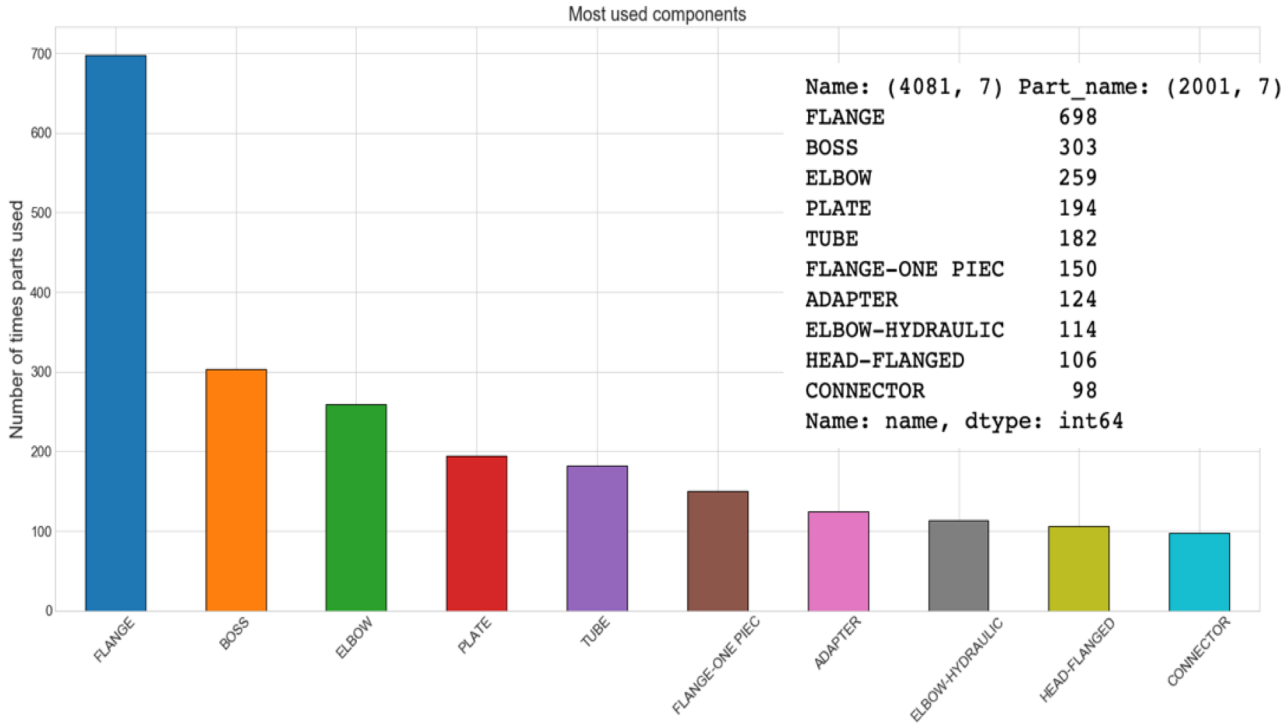
Cost Vs Tube Specifications by Orientation



- Higher bend radius, longer length oriented assemblies are costlier.
- Cost of oriented assemblies not impacted by weight and diameter.

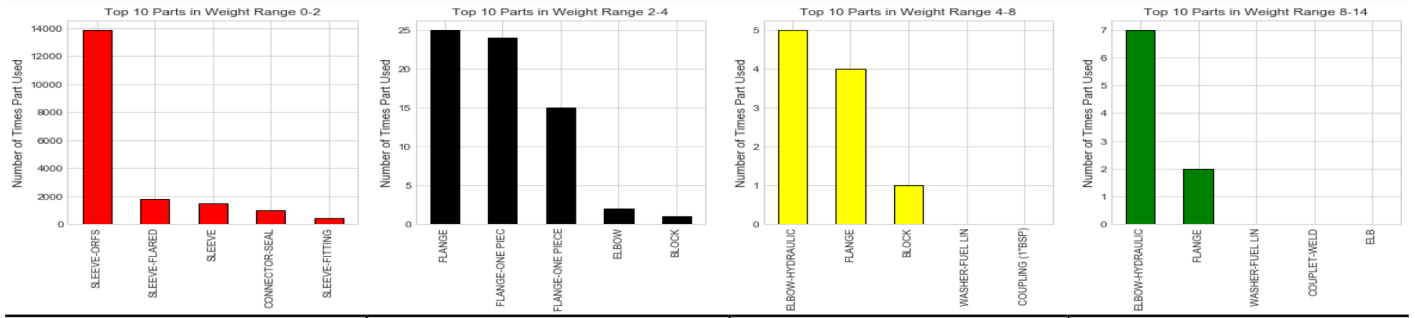
Most Used Assembly Components:

- Most used by overall usage



- Most used assembly weight

Weight 0-2	Weight 2-4	Weight 4-8	Weight 8-14



SLEEVE-ORFS	13851
SLEEVE-FLARED	1767
SLEEVE	1493
CONNECTOR-SEAL	957
SLEEVE-FITTING	455

FLANGE	25
FLANGE-ONE PIEC	24
FLANGE-ONE PIECE	15
ELBOW	2
BLOCK	1

ELBOW-HYDRAULIC	5
FLANGE	4
BLOCK	1

ELBOW-HYDRAULIC	7
FLANGE	2

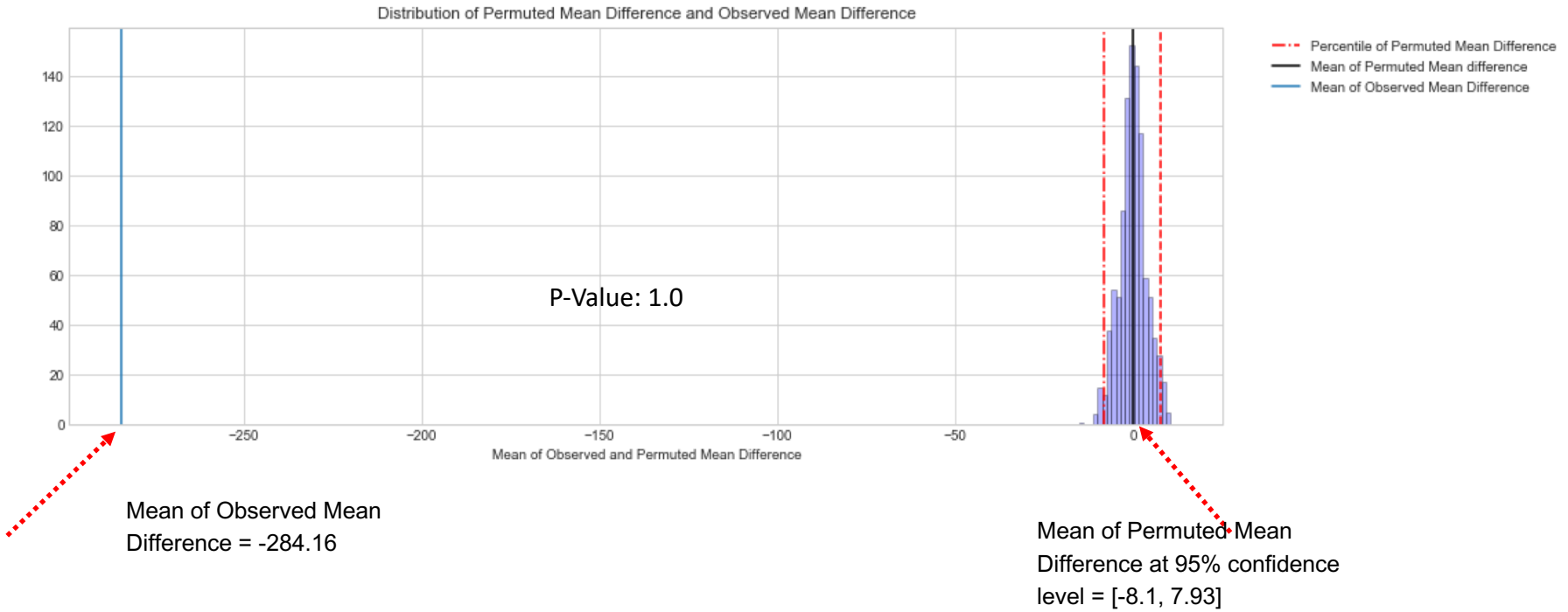
Bulk Vs Non-Bulk Assembly Statistical Inference

- Is bulk assembly different from non-bulk?

parameters	t-statistics	p_value
weight	-14.843691	1.360478e-49
annual_usage	-15.864254	2.373634e-56
min_order_quantity	-50.218847	0.000000e+00
quantity	18.082729	1.539973e-72
%supplier_rebate	82.783817	0.000000e+00
extended_cost	11.419592	4.092552e-30
total_cost	-11.578067	6.601705e-31
cost	-8.984690	2.807771e-19

Observation:
Yes, p-value < 0.05
Null rejected that assemblies are same.

- Is it possible that business stops buying assemblies in bulk?

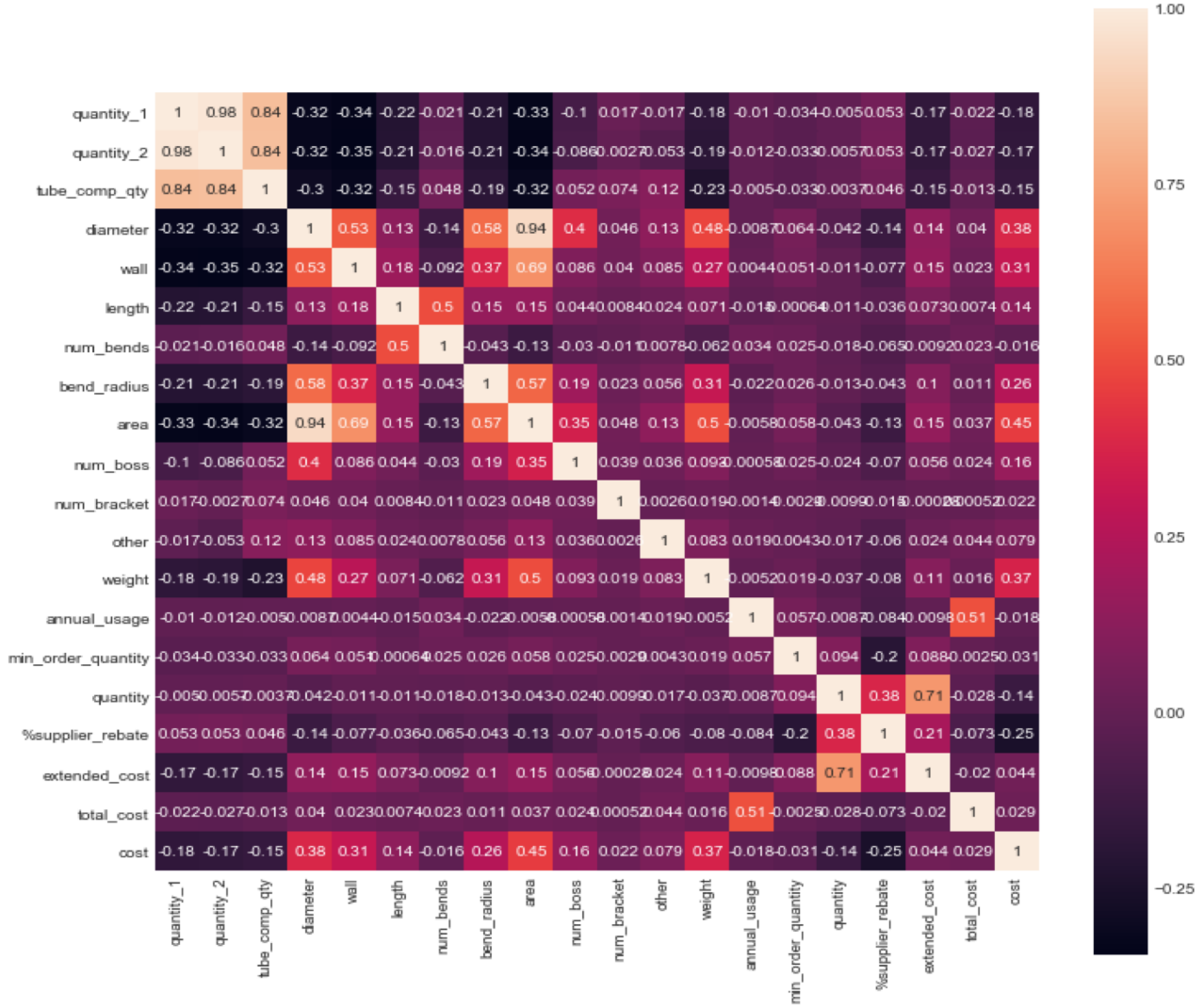


Observation:
Simulated Bootstrap sampling: 1000 examples
Significant difference between two mean values was -284.16 to 7.93
p-value > 0.05, Null accepted that bulk buying trend will continue.
Change would require significant variation in assembly parameters.

Check for Multi-Collinearity

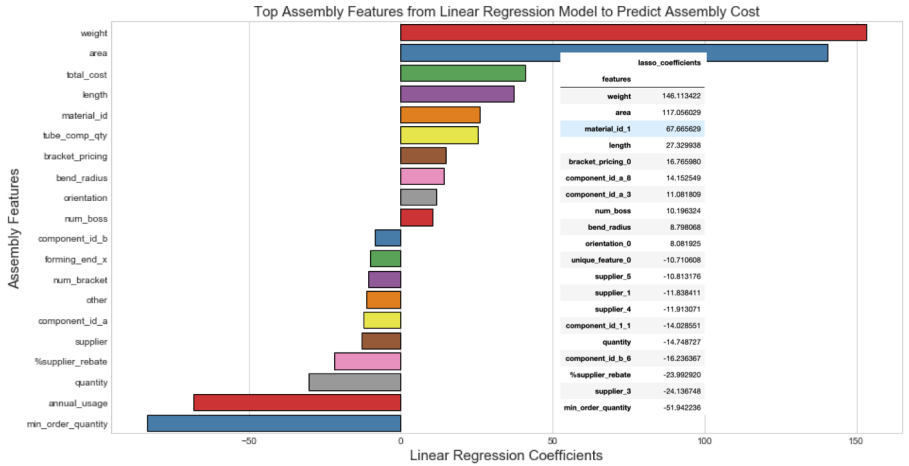
Collinear Features where Covariance is closer to 1, -1

- 'quantity_1', 'quantity_2' correlated to ‘tube_comp_qty’: covariance value of 0.84
- 'diameter' and 'wall' correlated to tube area: covariance value of 0.98 and 0.69
- ‘quantity’ and ‘extended cost’ correlated: covariance value of 0.71



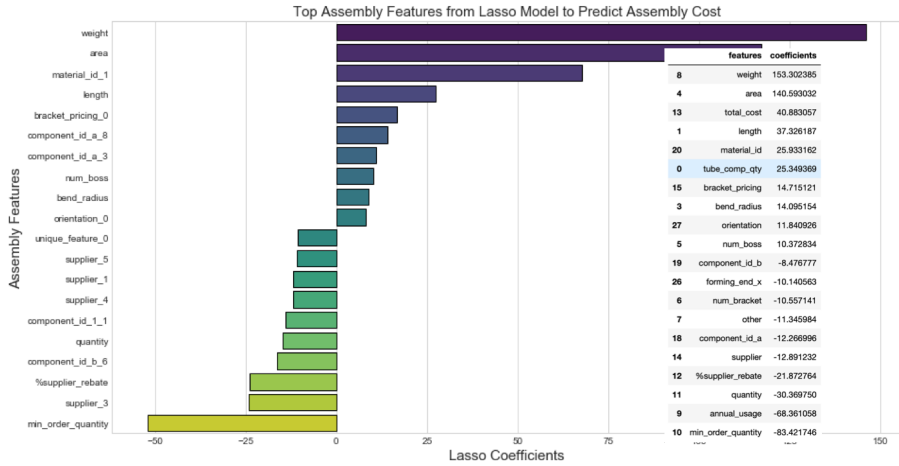
Feature Importance Comparison of Linear Model Vs Random Forest

Linear Regression: Top 20 features, Explained Variance 32%



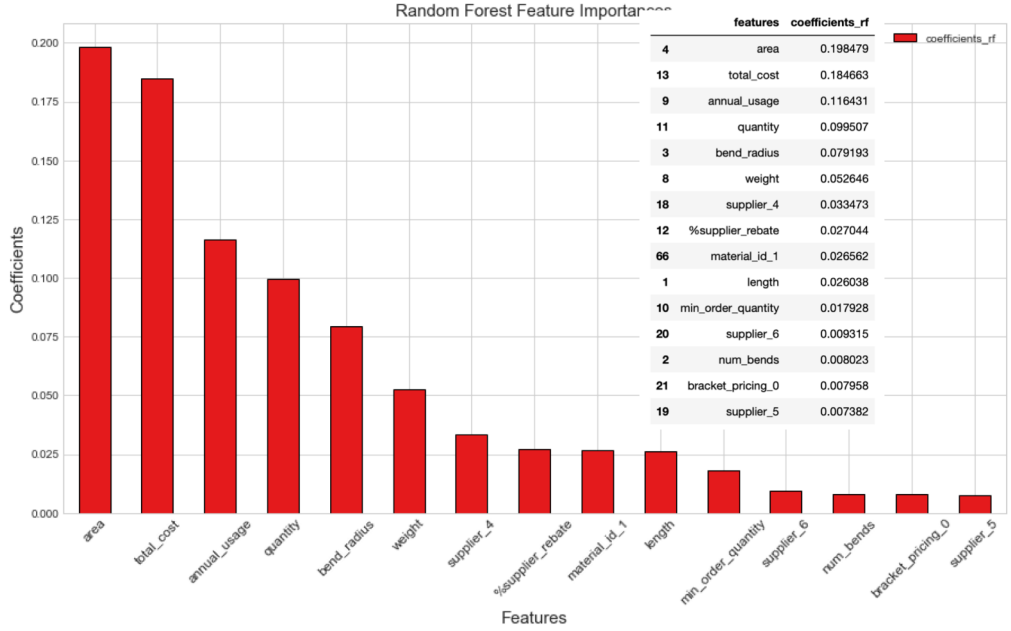
features	lasso_coefficients
weight	146.113422
area	117.056029
material_id_1	67.665629
length	27.329938
bracket_pricing_0	16.765980
component_id_a_8	14.152549
component_id_a_3	11.081809
num_boss	10.196324
bend_radius	8.798068
orientation_0	8.081925
unique_feature_0	-10.710608
supplier_5	-10.813176
supplier_1	-11.838411
supplier_4	-11.913071
component_id_1_1	-14.028551
quantity	-14.748727
component_id_b_6	-16.236367
%supplier_rebate	-23.992920
supplier_3	-24.136748
min_order_quantity	-51.942236

Lasso: Top 20 features, Explained Variance 42%



	features	coefficients
8	weight	153.302385
4	area	140.593032
13	total_cost	40.883057
1	length	37.326187
20	material_id	25.933162
0	tube_comp_qty	25.349369
15	bracket_pricing	14.715121
3	bend_radius	14.095154
27	orientation	11.840926
5	num_boss	10.372834
19	component_id_b	-8.476777
26	forming_end_x	-10.140563
6	num_bracket	-10.557141
7	other	-11.345984
18	component_id_a	-12.266996
14	supplier	-12.891232
12	%supplier_rebate	-21.872764
11	quantity	-30.369750
9	annual_usage	-68.361058
10	min_order_quantity	-83.421746

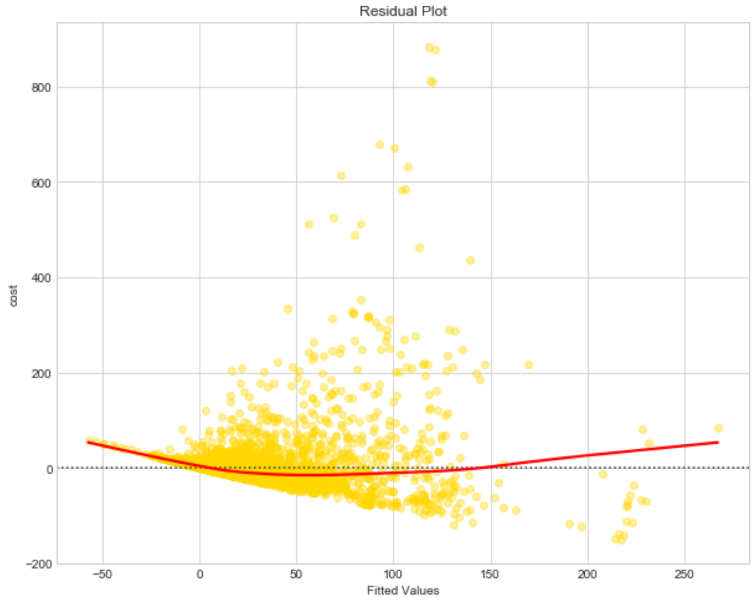
Random Forest: Top 15 features, Explained Variance 89%



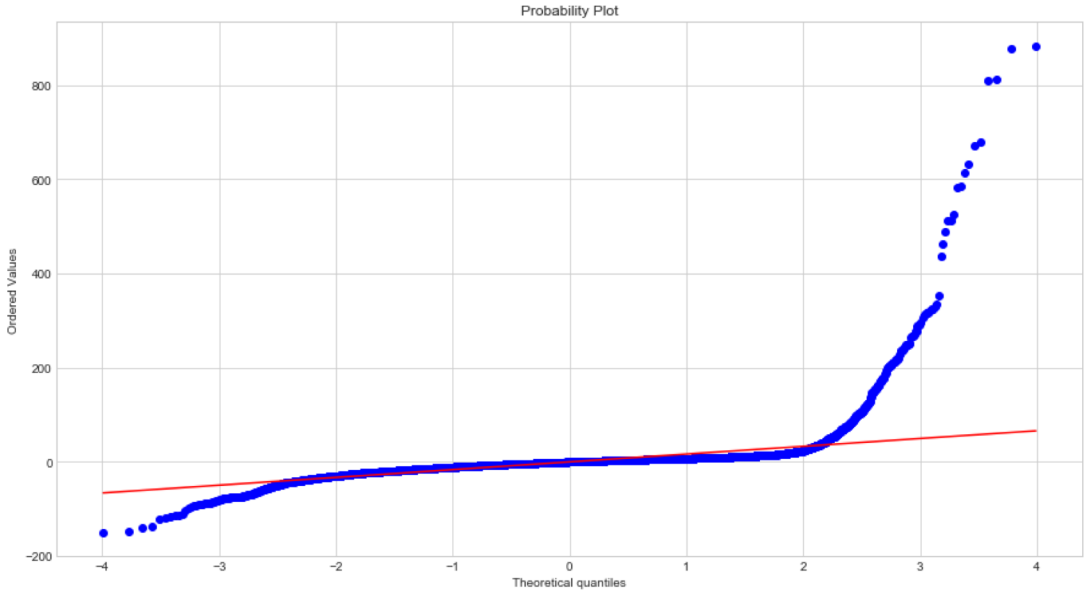
	features	coefficients_rf
4	area	0.198479
13	total_cost	0.184663
9	annual_usage	0.116431
11	quantity	0.099507
3	bend_radius	0.079193
8	weight	0.052646
18	supplier_4	0.033473
12	%supplier_rebate	0.027044
66	material_id_1	0.026562
1	length	0.026038
10	min_order_quantity	0.017928
20	supplier_6	0.009315
2	num_bends	0.008023
21	bracket_pricing_0	0.007958
19	supplier_5	0.007382

Reason for Lower Explained Variance in Linear Models

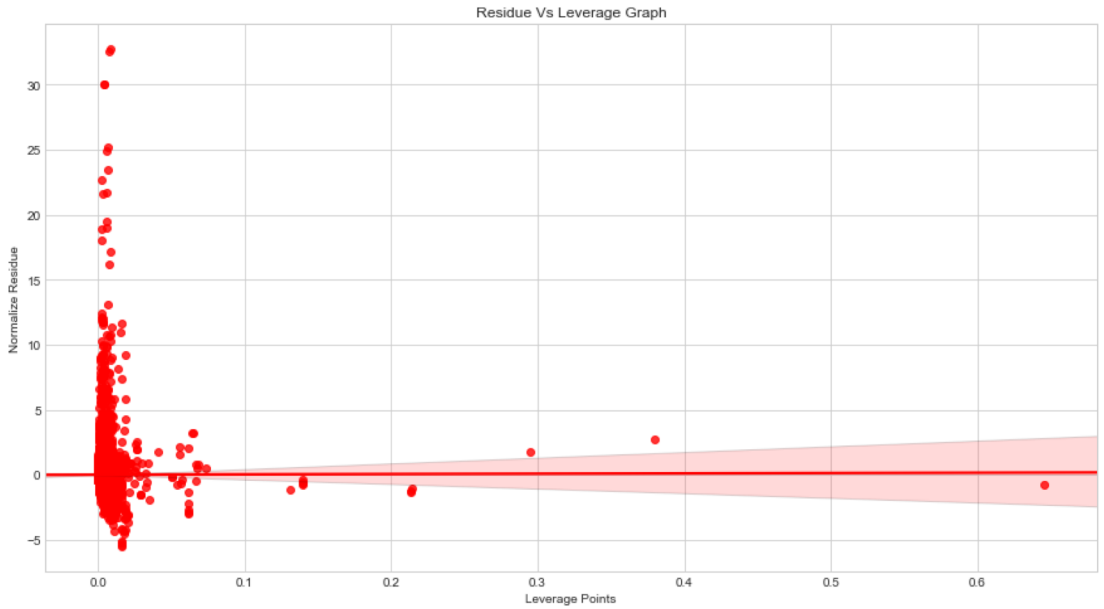
Skewness in Residual Distribution Plot



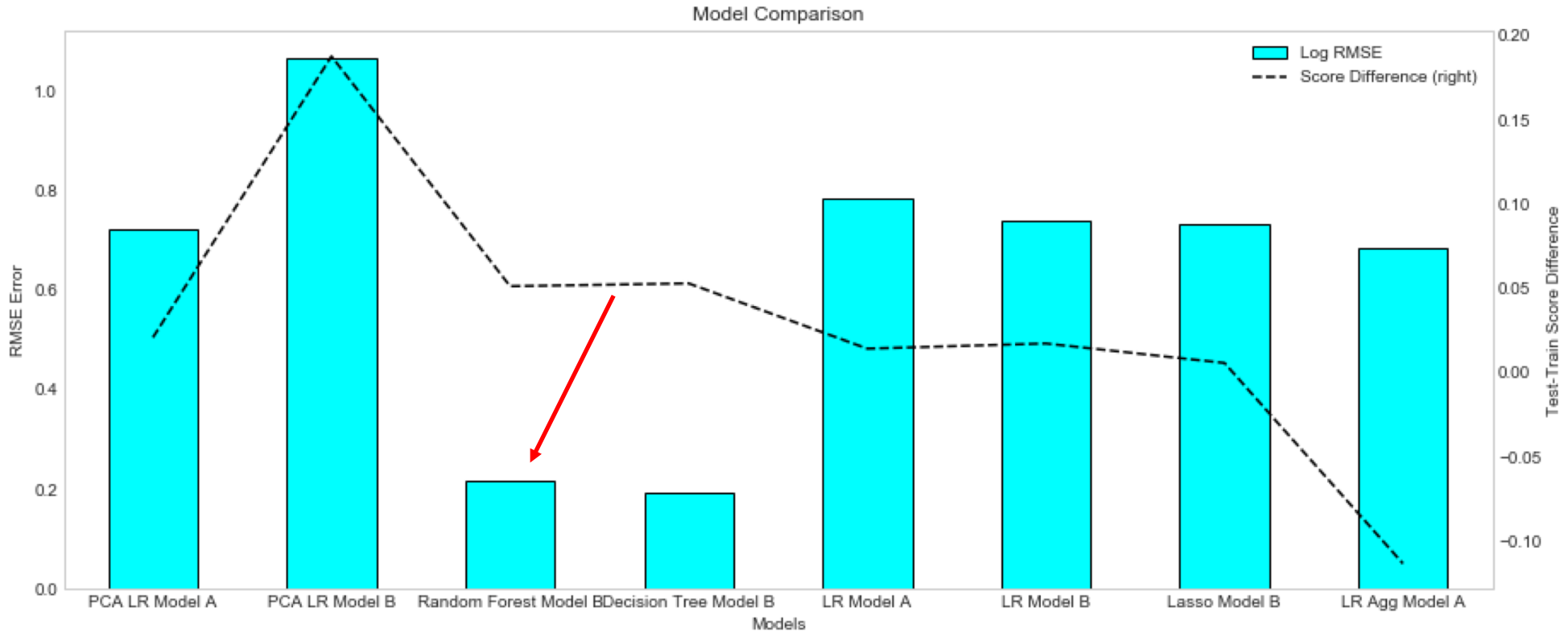
Outliers in Residual Values



Leverage Points in Residual Values



Supplier Price Prediction: Model Evaluation Summary



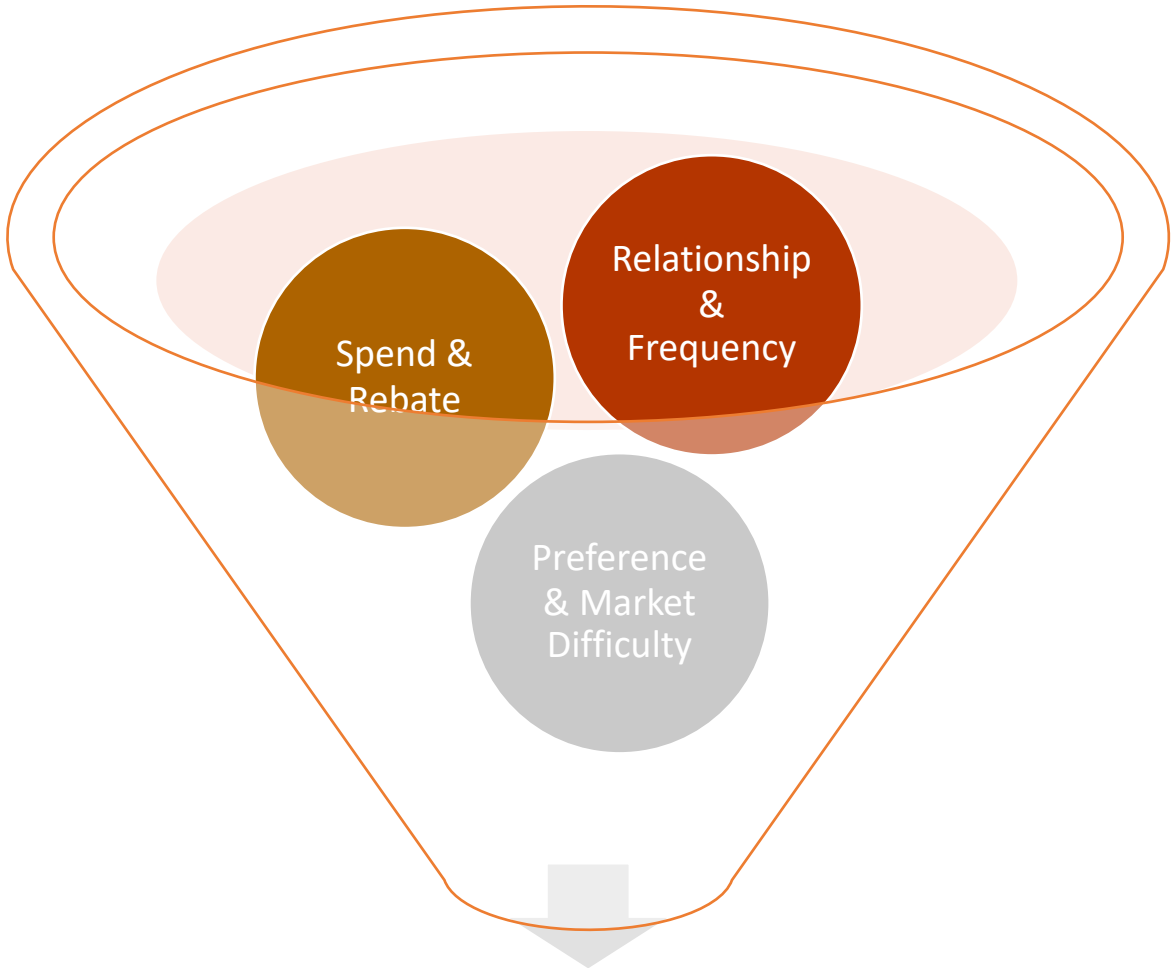
Observation:

1. Random Forest: Low log RMSE value of 0.21; Train-Test score difference of 5.1%.
2. Linear models: Comparable log RMSE values, except PCA LR on model B due to presence of categorical values.
3. **Random Forest, after hyperparameter tuning gave 98% accuracy; top 15 features with 89% explained variance.**

Criteria for Segmenting Assemblies and Suppliers

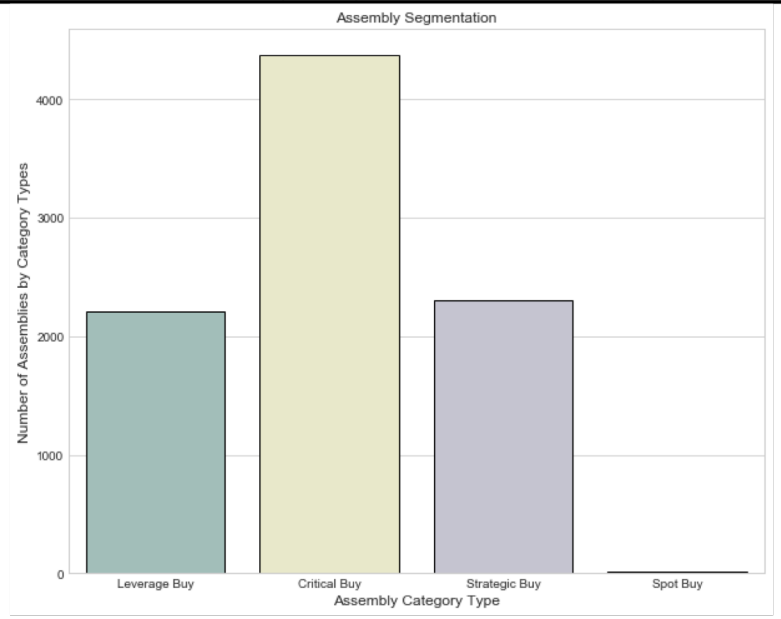
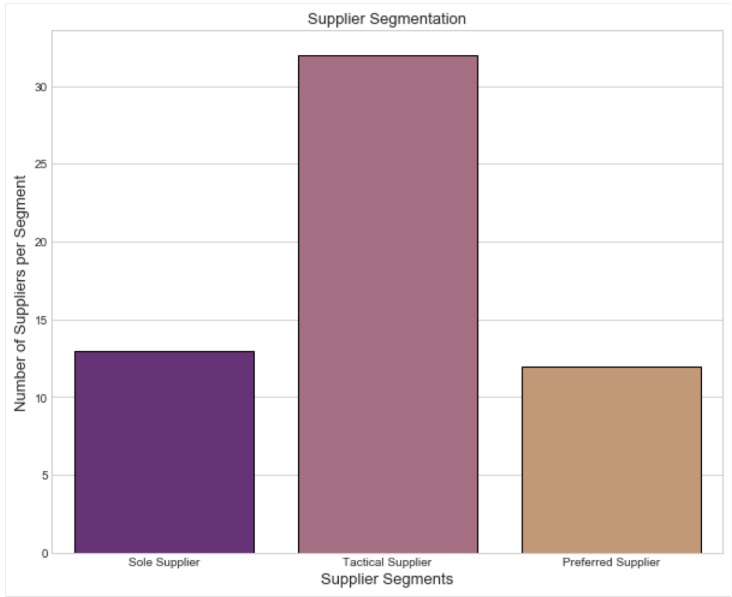
Business Needs:

- Spend: Total Spend Bulk/Non-Bulk
- Rebate: Discounts
- Relationship: Length of Contract
- Frequency: Number of Times Purchased
- Market Difficulty: Ease in switching supplier
- Preference: Supplier with maximum assortment
- Recency: Most Recent Purchase



Pre-Segmentation Breakdown of Assembly and Supplier Categories:

- Critical Assemblies and Tactical Suppliers make up 49% and 56.1% of the overall category breakdown.

Pre-Assembly Segmentation: 4 Categories	Pre-Supplier Segmentation: 3 Categories
	

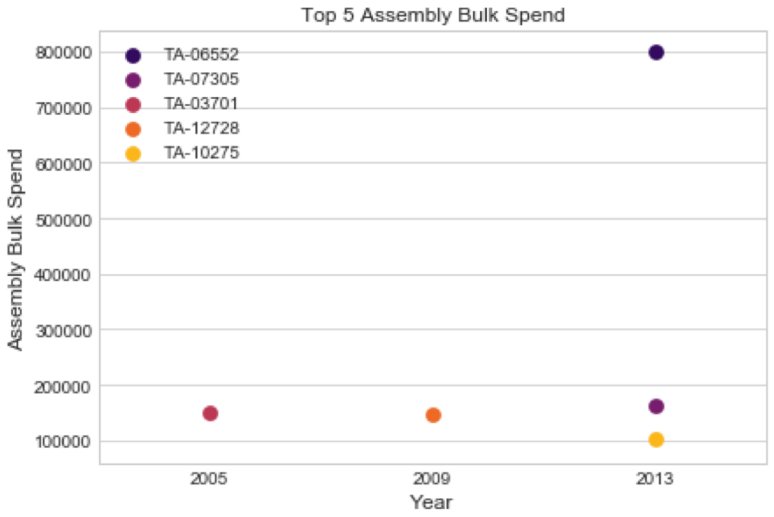
• Assembly Portfolio Breakdown:
  ○ Critical Buy 49.1%
  ○ Strategic Buy 25.9%
  ○ Leverage Buy 24.8%
  ○ Spot Buy Negligible
	• Supplier Portfolio Breakdown:
	  ○ Tacticle Suppliers 56.1%
	  ○ Sole Suppliers 22.8%
	  ○ Preferred Suppliers 21.1%

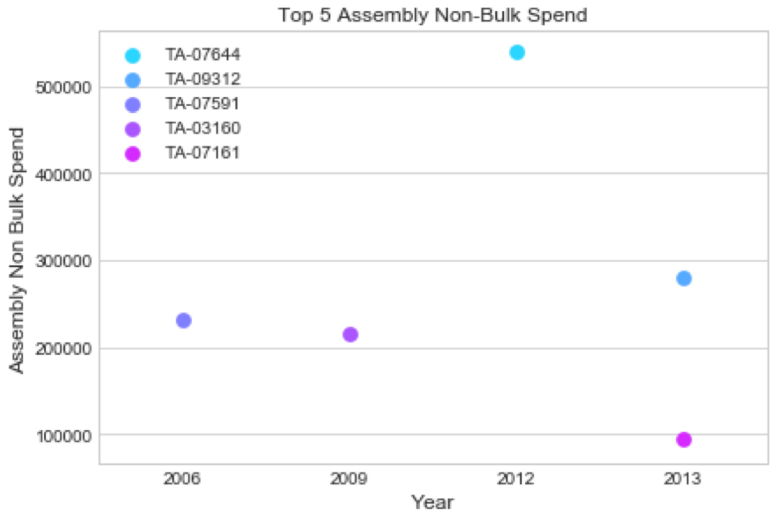
Top 5 Assemblies and Suppliers for Bulk and Non-Bulk Spend:

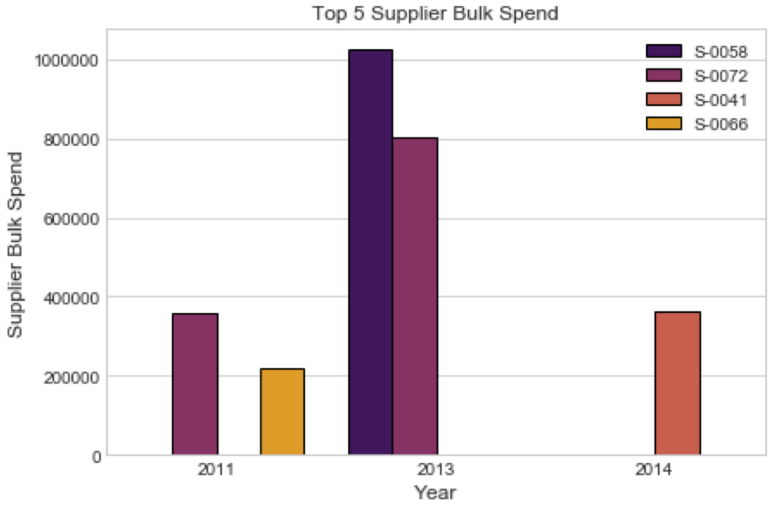
2005 2009 2013

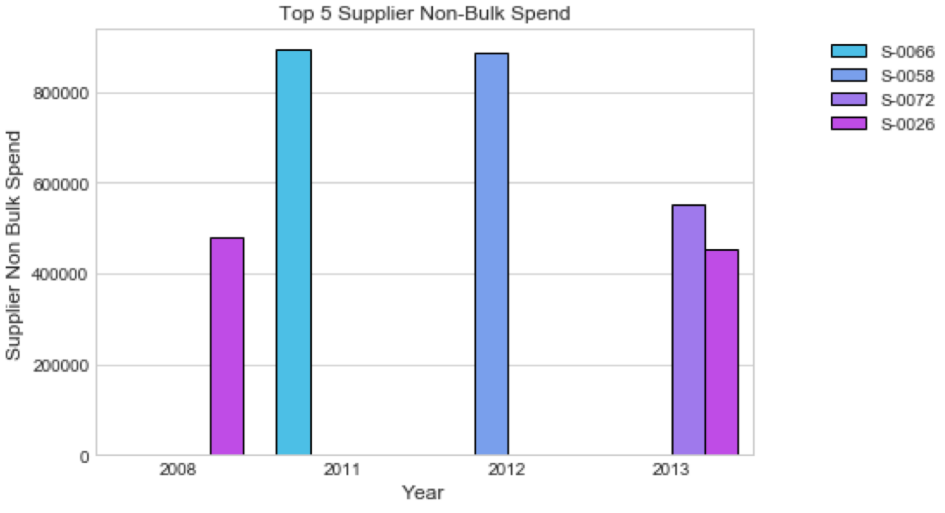
2006 2012 2014

Spend > 100K



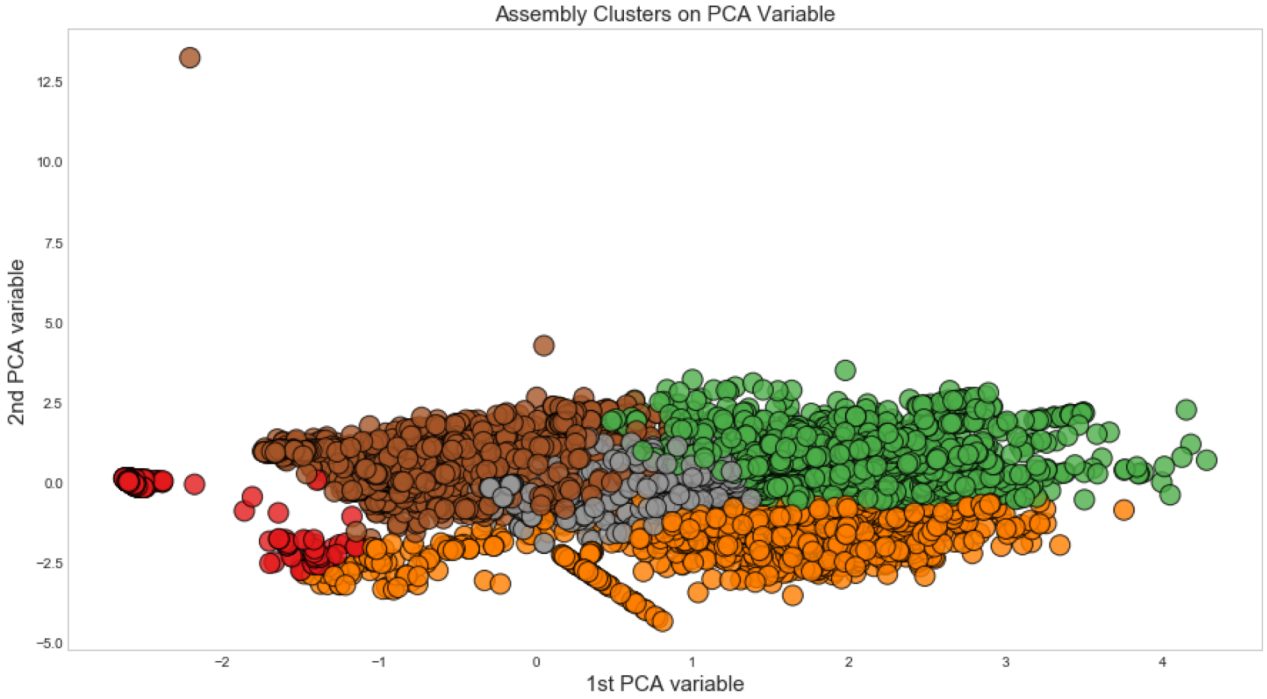




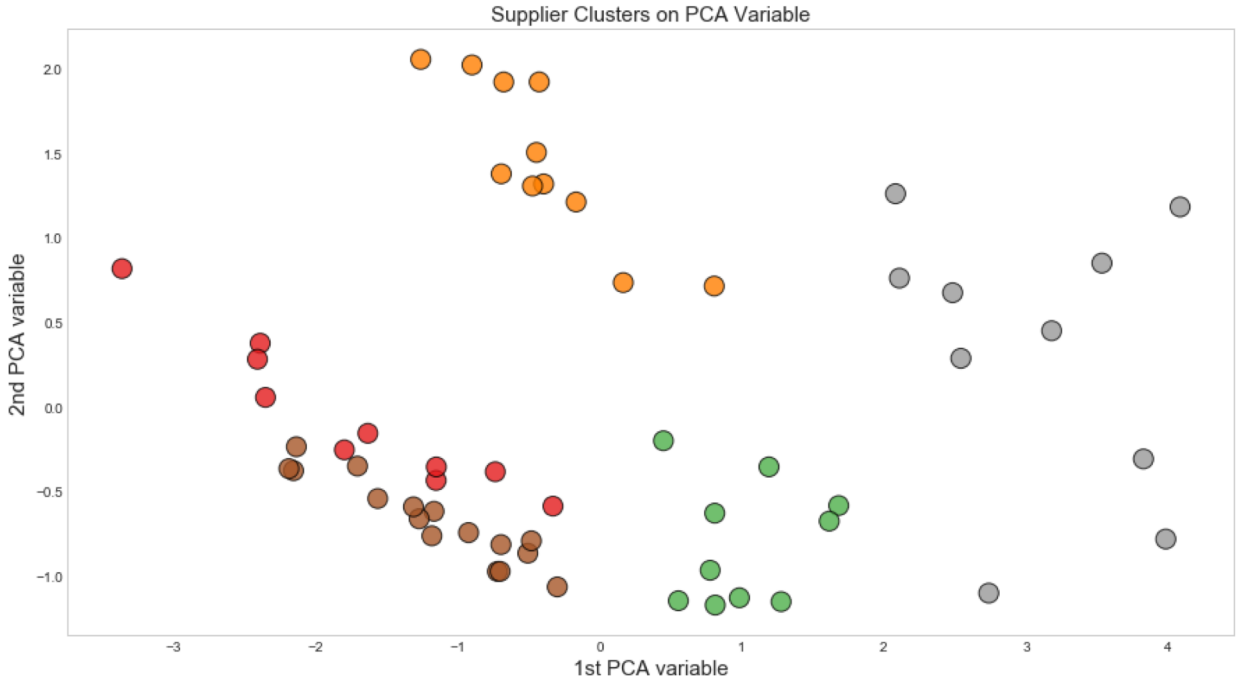


Clustering Quality

Number of Assembly Clusters = 5
Silhouette Score 0.46
Model: K-Means
Good Dense Formation

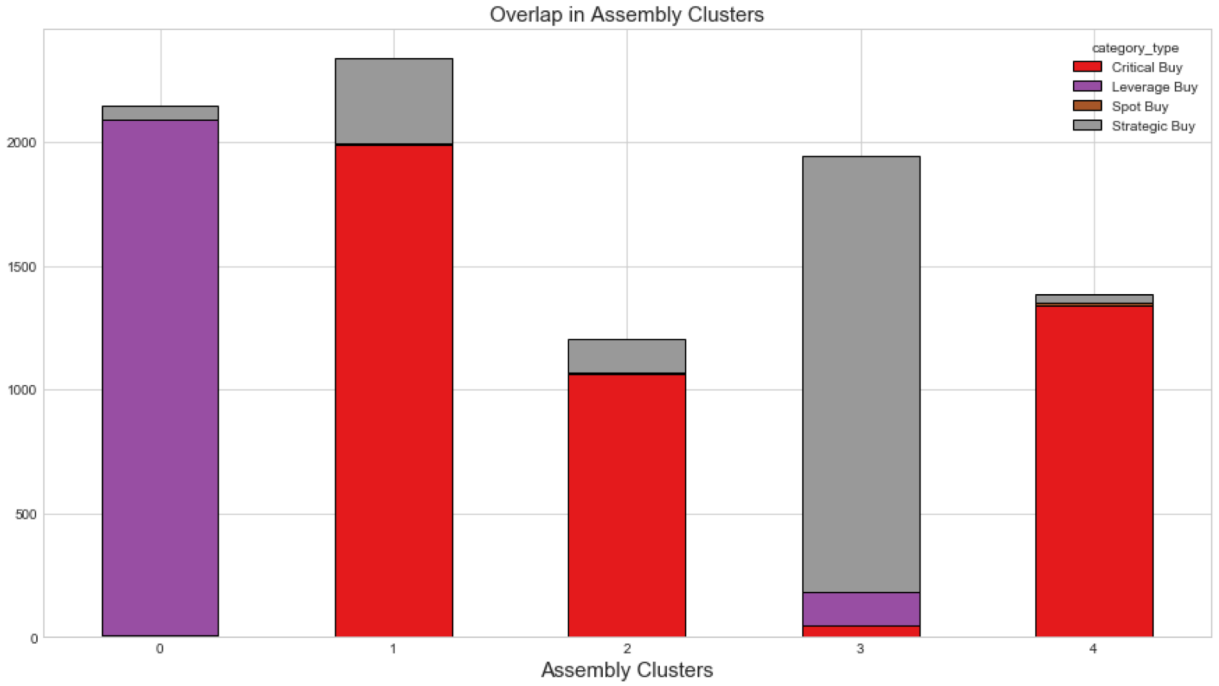


Number of Supplier Clusters = 5
Silhouette Score 0.35
Model: K-Means
Less Dense Formation



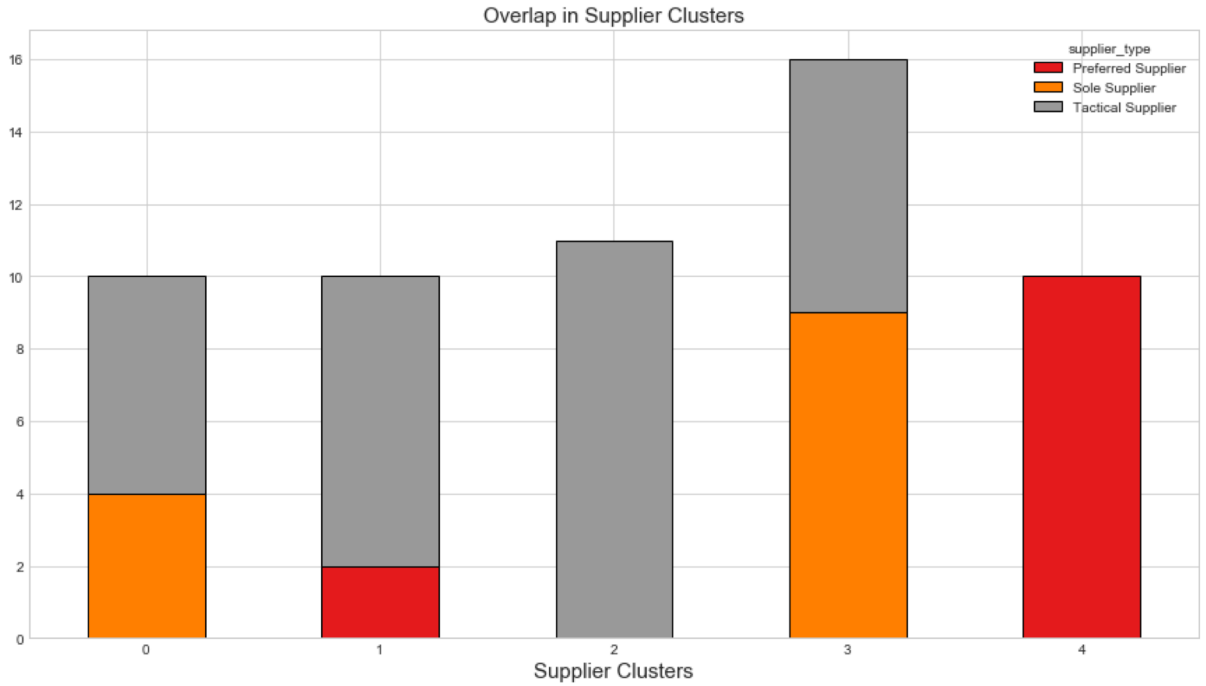
Clustering Overlap

All assembly categories have some strategic buy



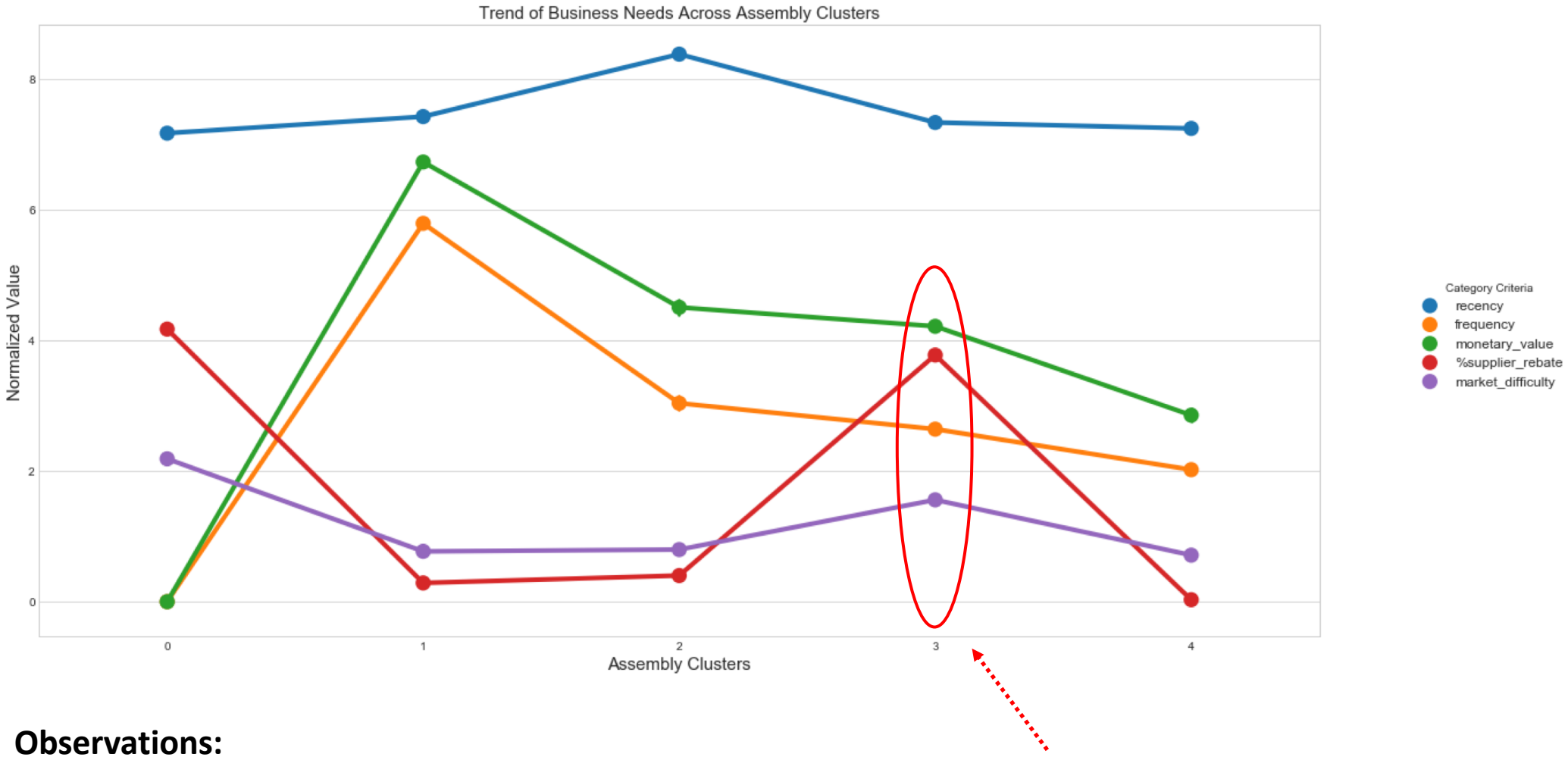
category_type	Critical Buy	Leverage Buy	Spot Buy	Strategic Buy
assembly_clusters				
0	6	2083	0	58
1	1990	0	5	343
2	1063	5	1	135
3	48	136	0	1758
4	1339	1	11	32

Tactical suppliers have some sole and preferred suppliers



supplier_type	Preferred Supplier	Sole Supplier	Tactical Supplier
supplier_clusters			
0	0	4	6
1	2	0	8
2	0	0	11
3	0	9	7
4	10	0	0

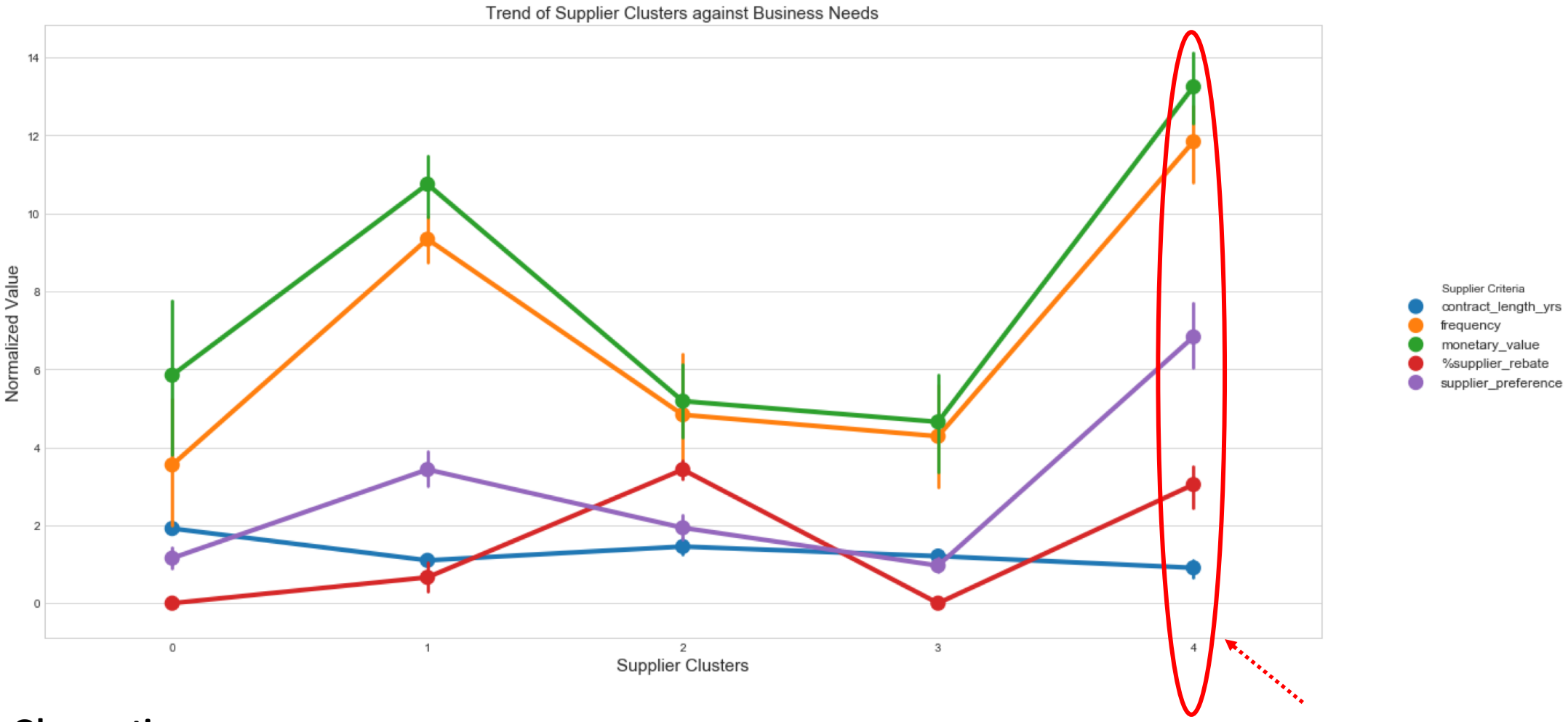
Best Managed Assembly Cluster #3



Observations:

- Clusters 1,2 and 4: High supplier spend, high frequency purchase, lowest supplier rebates.
- Cluster 0: No spend, offering offer high rebates (potential for leverage buy).
- **Cluster 3 (Best Managed): High rebates, fewer suppliers managing high spend (like strategic buy).**

Best Managed Supplier Cluster #4



Observations:

- Clusters 0, 3: Lowest supplier preference, no rebates (similar to Sole Suppliers). Opportunity for spend consolidation.
- Cluster 2: Similar to cluster 4 with very low supplier preference. Opportunity to develop cluster 2.
- Cluster 1: High spend new preferred suppliers with relatively low rebate. Opportunity for contract negotiation.
- **Cluster 4 (Best managed): High rebates from highly preferred suppliers (similar to Preferred Suppliers).**

Project Conclusion

- By using Random Forest and K-Means, we were able to establish pricing prediction and benchmark assemblies and supplier clusters from a pool of 8,855 unique assemblies and 57 suppliers.

- Out of multiple assembly features and observations, we were able to select out top 15 assembly parameters (both linear and non-linear) that explained 89% variance in the predicted supplier price with 98% accuracy.

- By choosing optimal k-value, cross tab count and silhouette score, we were able to evaluate quality of grouped clusters and identify key business trends and opportunities.

Future Possibilities

- Monitor Supplier Performance: We can add other features such as on-time delivery, safety statistics, contract compliance and inventory levels to gain further supplier insights.

- Build Real Time Dashboards: We can connect our model to real time data feed to collect actionable insights on the fly.

- Further Modeling: We can try other algorithms such as time-series to predict annual supplier pricing.

Thank You